Principles of Responsible Work

by James Bach, Jon Bach, and Michael Bolton, v2.1, 9/7/26

1. **Every non-trivial business comprises some set of services that enable it to function.** Examples include sales, accounting, R&D, customer support, etc. These services must be sufficiently reliable or else the business will collapse.

2. **Every service entails the risk of failure.** When failures occur, the business must be able to recognize them and recover. In regulated industries, risk management may be subject to specific process mandates.

3. **A "responsible person" is a natural person in a business who is reasonably competent, prepared, and accountable for some service that sustains or defines that business.** No matter what tools or processes are used within a business, someone must be responsible for them. To bear responsibility, a person must have sufficient capacity. For instance, neither a child nor a tool (such as AI) has the capacity (either legally or socially) to bear responsibility. Even adult humans may lack capacity, such as when an airline pilot has had insufficient sleep or is under the influence of drugs.

4. **A "responsible service" is one that is performed in good faith by a responsible person.** This may include interpreting and following procedures, improving skills, anticipating problems, and reporting to relevant authorities or clients, both inside and outside the business.

5. **Responsible services may incorporate any manner of tool, as long as the person performing that service can operate the tool safely, legally, and with reasonable efficiency.** The effort and skill required to cost effectively operate a tool generally increases with the cost of the tool, the complexity of the tool, the obscurity of its output, the amount of output produced per unit of time, and the reliability of the tool when performing that task. The operator must also reasonably anticipate outages and breakdowns of their tools.

6. **Responsibility can be taken, shared, declined, or delegated, as long as there is a clear and reasonable protocol for doing so.** In the absence of such a protocol, the business is vulnerable to accusations of negligence or breach of contract. This is a principal topic of common law and the law of contracts, although specific laws and regulations may constrain how a business can distribute responsibility. Even if a business provides a service "as is" they may have a strong competitive incentive to maximize quality or trust via the mechanism of responsibility.

7. **Therefore, to avoid inefficiency, poor quality, and legal trouble, businesses must develop and maintain clear lines of responsibility, assure competence and readiness among responsible persons, put reliable tools in place, and maintain appropriate oversight of any delegated responsibility.**

Responsible Operation of AI

1. **AI cannot bear responsibility.**[1] AI is not a responsible person, and it would be meaningless to speak of a tool that operates in "good faith." Therefore, it cannot provide a responsible service, nor can responsibility be delegated to it.
2. **An "AI agent" is always a tool operated by a natural person,** irrespective of whether the person is monitoring it in real-time.
3. **Thus, the operator of an AI agent always bears responsibility for the behavior of that agent.**[2] This includes anticipating availability issues, such as outages, token rationing, or poor performance.
4. **The responsible operator must assure adequate quality of the work**; they cannot merely prompt and pray.

Therefore, the operator must...

5. **be sufficiently skilled in the use of the AI tool.**
6. **be sufficiently prepared to operate the tool in that context.**
7. **be sufficiently alert to risks, anomalies, or defects that may occur in the work.**[3]
8. **reasonably anticipate restrictions or interruptions of the services on which the work depends.**
9. **feel empowered (and actually have the power) to reject or remediate any work done by AI.** Otherwise, the operator becomes a scapegoat, a "moral crumple zone."[4]
10. **avoid or mitigate the special hazards of AI operation.** See below.

---

[1] Wein, L. E. (1992). Responsibility of intelligent artifacts: Toward an automation jurisprudence. *Harv. JL & Tech.*, *6*, 103. "Currently the law does not allow us to sue a machine, although it seems that some machines are beyond our control. We assign liability exclusively to humans, not to quadrupedal or inanimate entities. If someone is run over by an automobile, we do not seek to destroy the car, but instead we penalize some human we hold responsible for the mischief."

[2] There are relevant concepts in law that make employers liable even for their *human* agents, such as *vicarious liability* and *apparent authority.* But, as a mere machine, AI does not have any societal standing and *cannot* be held liable for its actions.

[3] Klein, G., Pliske, R., Crandall, B., & Woods, D. D. (2005). *Problem Detection.* Cognition, Technology & Work, *7*(1), 14–28. https://doi.org/10.1007/s10111-004-0166-y

[4] See explanation below.

Special Hazards of AI Operation

Technological

- **Service Outage.**[5] Online AI services are subject to outages. Locally hosted AI can mitigate this risk, but frontier models are generally not locally available. Even in the absence of a hard outage, a service may experience data center failures or excessive demand that forces them to ration tokens. The operator of the AI may not have a viable alternative for performing the service.
- **Service Adulteration.** This includes any intermittent or progressive degradation of the AI service used by the operator relative to their reasonable expectation, perhaps due to factors such as model drift,[6] context drift,[7] model poisoning,[8] model collapse,[9] or inadequate performance.

Interactional

- **Cognitive Overload.**[10] This occurs when the AI produces too much output, the output is too complex, or it is produced too quickly for the operator to process it. Overload can lead to poor supervision of AI, allowing mistakes to go unnoticed. It creates chronic stress, threatening burnout. It leads directly to cognitive debt, and also encourages cognitive surrender.
- **Cognitive Debt.**[11] When knowledge workers do their own work, they learn as they go. By the time they deliver it, they know quite a lot about their material and are able to answer questions and challenges. But AI can produce results much faster than a human operator can study them. When a person falls substantially behind the AI, that is "cognitive debt." With a large enough backlog of learning not completed, their ability to supervise the AI can collapse. They then cease to feel that "their" work belongs to them, even though no one else is in a position to be accountable for that work.

---

[5] Chu, X., Talluri, S., Lu, Q., & Iosup, A. (2025). An Empirical Characterization of Outages and Incidents in Public Services for Large Language Models. *Proceedings of the 16th ACM/SPEC International Conference on Performance Engineering*, 69–80. https://doi.org/10.1145/3676151.3719372

[6] Surya Gangadhar Patchipala. (2023). Tackling data and model drift in AI: Strategies for maintaining accuracy during ML model inference. *International Journal of Science and Research Archive*, *10*(2), 1198–1209. https://doi.org/10.30574/ijsra.2023.10.2.0855

[7] Dongre, V., Rossi, R. A., Lai, V. D., Yoon, D. S., Hakkani-Tür, D., & Bui, T. (2025). *Drift No More? Context Equilibria in Multi-Turn LLM Interactions* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2510.07777

[8] Tian, Z., Cui, L., Liang, J., & Yu, S. (2023). A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. *ACM Computing Surveys*, *55*(8), 1–35. https://doi.org/10.1145/3551636

[9] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

[10] Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2506.08872

[11] Ibid.

- **Cognitive Atrophy.**[12,13] This occurs when an operator of AI who would otherwise be competent to supervise it progressively loses their ability to judge the work, because the person stops getting regular practice and experience. Knowledge that they would otherwise quickly bring to mind grows stale and recedes from their awareness.

- **Cognitive Surrender.**[14,15] This is when the operators of AI cease thinking critically about the work produced by AI. It is the passive acceptance of whatever AI gives them. In any situation where a human is expected to take responsibility for the work, it amounts to a complete abdication of that responsibility.

- **Anthropomorphism and Anthropomorphizing.**[16] This means treating AI as if it were human, i.e. having emotions, empathy, agency, and associated rights and privileges that humans acquire by right of birth and by participating in society. AI does not have these things. Even when AI behaves in ways that are consistent with that of a reasonable adult, it does so for reasons that have nothing to do with human motivations or human experience, nor is its behavior in one context predictive of its behavior in other contexts. Anthropomorphism leads to unhealthy parasocial relationships and unjustified trust.

- **Misalignment.**[17] AI does not necessarily behave in accordance with the intentions and values of the people who operate it. An agent may misbehave in shocking ways, including cheating, lying, or doing other things that we would consider malicious or diabolical if a human were to do the same things.

- **Automation Bias.**[18] This is when people trust the output and behavior of machines more than their own judgment, even when faced with clear evidence that the machine is incorrect or malfunctioning.

- **Chronic Stress.**[19,20] Whether supervising, training, testing, or collaborating with AI, stress can become chronic. This can lead to burnout, excessive turnover, or counterproductive behavior.

---

[12] Ginac, F. (2026). *Cognitive Atrophy and Systemic Collapse in AI-Dependent Software Engineering* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2604.26855

[13] Bainbridge, L. (1983). Ironies of Automation. In *Analysis, Design and Evaluation of Man–Machine Systems* (pp. 129–135). Elsevier. https://doi.org/10.1016/B978-0-08-029348-6.50026-9

[14] Shaw, S. D., & Nave, G. (2026). *Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender*. PsyArXiv. https://doi.org/10.31234/osf.io/yk25n_v1

[15] *"So, You 10x'd Your Work",* https://www.satisfice.com/blog/archives/488009

[16] Reani, M., He, X., Luo, Y., & Sun, Z. (2025). *Fundamental Over-Attribution Error: Anthropomorphic Design of Ai and its Negative Effect on Human Perception*. SSRN. https://doi.org/10.2139/ssrn.5222775

[17] Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Vierling, L., Hong, D., Zhou, J., Zhang, Z., Zeng, F., Dai, J., Pan, X., Ng, K. Y., O'Gara, A., Xu, H., Tse, B., … Gao, W. (2025). *AI Alignment: A Comprehensive Survey* (arXiv:2310.19852). arXiv. https://doi.org/10.48550/arXiv.2310.19852

[18] Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In *Decision making in aviation* (pp. 289–294). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315095080-17/automation-bias-intelligent-time-critical-decision-support-systems-cummings

[19] Hou, Y., & Fan, L. (2024). Working with AI: The Effect of Job Stress on Hotel Employees' Work Engagement. *Behavioral Sciences*, *14*(11), 1076. https://doi.org/10.3390/bs14111076

[20] Kim, B.-J., Oh, H.-J., Kim, M.-J., & Lee, D. (2024). The Perils of Perfection: Navigating the Ripple Effects of Organizational Perfectionism on Employee Misbehavior through Job Insecurity and the Buffering Role of AI Learning Self-Efficacy. *Behavioral Sciences*, *14*(10), 937. https://doi.org/10.3390/bs14100937

Managerial

- **Data Negligence.** As opposed to data governance,[21] this covers any misuse of data in the training or operation of AI, or the insufficiency of systems and protocols for handling and protecting data.
- **Reckless Spending.**[22] AI provided as a service is usually not a fixed and predictable cost. It depends on usage and the particulars of the model. An AI operator may bust their budget with careless or otherwise excessive use.
- **Violations of Law.** Depending on location,[23] there may be laws relating to the use or abuse of AI technology. Operators of AI must make themselves aware of those laws.
- **Moral Crumple Zone.**[24,25,26] This can occur in an automated system that has a human operator or supervisor (such as a self-driving car with a safety driver, or an ordinary user of ChatGPT). It happens when a failure of the *system* is routinely and carelessly blamed on the *human*. The human functions as a sort of "crumple zone" in the moral sense: a component designed to assume blame in order to deflect it from the automation.
- **Business Disruption.**[27,28] While any new technology will be somewhat disruptive, AI is special, because it is *specifically designed* to simulate human agency and blur the lines between machines and people. There are potentially far-reaching and profound effects on processes and personnel of introducing AI into an organization. These include loss of tacit and tribal knowledge, devaluation of craftsmanship, demoralization, alienation, perverse incentives, and disruption of the human experiential learning pipeline.

---

[21] Pahune, S., Akhtar, Z., Mandapati, V., & Siddique, K. (2025). The Importance of AI Data Governance in Large Language Models. *Big Data and Cognitive Computing*, *9*(6), 147. https://doi.org/10.3390/bdcc9060147

[22] Bai, L., Huang, Z., Wang, X., Sun, J., Mihalcea, R., Brynjolfsson, E., Pentland, A., & Pei, J. (2026). *How Do AI Agents Spend Your Money? Analyzing and Predicting Token Consumption in Agentic Coding Tasks* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2604.22750

[23] https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[24] Elish, M. C. (2019). *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction.* Engaging Science, Technology, and Society, *5*, 40–60. https://doi.org/10.17351/ests2019.260

[25] The Open AI Service Agreement includes a maximal disclaimer that essentially says its users operate the product at their own risk. https://openai.com/policies/services-agreement/

[26] Green, B. (2022). The Flaws of Policies Requiring Human Oversight of Government Algorithms. *Computer Law & Security Review*, *45*, 105681. https://doi.org/10.1016/j.clsr.2022.105681

[27] Hu, Q., Xiao, Q., Cao, H., & Shen, H. (2026). When Your Boss Is an AI Bot: Exploring Opportunities and Risks of Manager Clone Agents in the Future Workplace. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–21. https://doi.org/10.1145/3772318.3790987

[28] Lee, C. P., Lee, M. K., & Mutlu, B. (2026). *Making the Invisible Visible: Understanding the Mismatch Between Organizational Goals and Worker Experiences in AI Adoption* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2605.03078

Discussion

The Quality of AI is Not the Central Issue; The Central Issue is Responsibility

At the time of writing, in 2026, the world is in a "wild west" era of AI adoption. It's a gold rush. Responsible use of AI is lagging far behind reckless use. Hence, this document. We are working toward a societal consensus about responsible work and responsible use of AI. The thrust of our argument is that human oversight and "human-in-the-loop" are not merely techniques for dealing with unreliable AI, they are foundational requirements for ethical business practice *even with highly reliable AI*.

The widespread intrusion of GenAI technology into social workspaces brings new urgency to the problem of responsible work. GenAI may credibly simulate human intellectual performance. Indeed, there are many who wonder: why shouldn't society embrace AI tools that produce work good enough to satisfy casual or even expert review?

Notice that this frames the matter as an issue of quality. But we think that is the wrong framing. It's kind of like arguing that we should all prefer to be governed by an absolute monarchy as long as the cost of living is low and most of us are enjoying ourselves.

Even the quality argument is weak. Yes, GenAI *can* produce excellent work. But *can* is not *will*. The Space Shuttle was once a craft that *could* be used to put satellites into orbit. It doesn't do that anymore because the risks and costs were found to outweigh the benefits. Similarly, we cannot rationally consider the benefits of GenAI without also investigating its failures.[29] This requires extensive testing and analysis, which in the case of GenAI is often prohibitively expensive. Apart from the problem of testing, there is the problem of hastiness. GenAI is being widely adopted, at the time we write this, in an almost frantic atmosphere that discourages anyone from looking too closely at what they are doing.

A better frame of reference than quality is responsibility: *when* there is a problem, *whose* problem is it? Who will make it right? Who bears the burden of risk? As a tool, AI cannot be responsible for itself. It cannot participate in the mechanisms of human society that construct responsibility. This leads to moral hazard,[30] which refers to a tendency to behave recklessly whenever one is insulated from the consequences of one's behavior. One form of moral hazard is due to GenAI being designed to operate in the social world, as if it were a creature with agency and standing—yet it cannot and will not suffer any consequences from its behavior. It is unconstrained by any rational concern for its own well-being. Another form of moral hazard

---

[29] The AI Incident Database is a useful resource for browsing the variety of failures that may occur. (https://incidentdatabase.ai/apps/incidents)

[30] Rowell, D., & Connelly, L. B. (2012). A History of the Term "Moral Hazard." *Journal of Risk and Insurance*, *79*(4), 1051–1075. https://doi.org/10.1111/j.1539-6975.2011.01448.x

occurs when humans hide behind AI and use it as an accountability sink.[31] For instance, an AI system for deciding who gets a home loan may behave in grossly unfair ways, yet its owners might successfully plead that they are shocked, *shocked*, to discover the bias.

A third source of moral hazard is the so-called AI Productivity Paradox:[32]

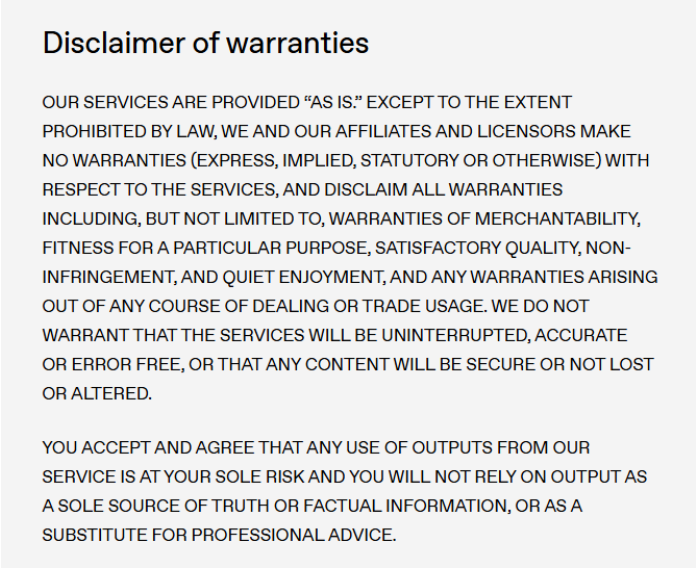
> *For many tasks, it takes considerable time and effort to use AI responsibly. Yet, for AI to be productive, it must save time and effort. Therefore, to maximize perceived AI productivity, we will be rewarded for using AI irresponsibly.*
>
> *In other words, the more desperately management wants the magic box to work, the less they will tolerate people like you looking "behind the curtain" of the magic [to make sure that it does work]! This paradox is the heart of our concern about AI.*

This paradox creates moral hazard through the mechanism of allowing both managers (who say "you did bad work") and workers (who say, "you rushed us") to point at the other party as being the root cause of any harm done by the AI, while excusing themselves from blame.

Regardless of how good a technology is, people operating within civil society cannot ignore the question of what happens when that technology fails or otherwise hurts someone.

It Can Be Okay to Use or Provide an Irresponsible Service

In our terms, ChatGPT is an example of an irresponsible service. That is to say, OpenAI expressly admits that its product is not necessarily safe and that you use it at your own risk.[33]



Disclaimer of warranties

OUR SERVICES ARE PROVIDED "AS IS." EXCEPT TO THE EXTENT PROHIBITED BY LAW, WE AND OUR AFFILIATES AND LICENSORS MAKE NO WARRANTIES (EXPRESS, IMPLIED, STATUTORY OR OTHERWISE) WITH RESPECT TO THE SERVICES, AND DISCLAIM ALL WARRANTIES INCLUDING, BUT NOT LIMITED TO, WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, SATISFACTORY QUALITY, NON-INFRINGEMENT, AND QUIET ENJOYMENT, AND ANY WARRANTIES ARISING OUT OF ANY COURSE OF DEALING OR TRADE USAGE. WE DO NOT WARRANT THAT THE SERVICES WILL BE UNINTERRUPTED, ACCURATE OR ERROR FREE, OR THAT ANY CONTENT WILL BE SECURE OR NOT LOST OR ALTERED.

YOU ACCEPT AND AGREE THAT ANY USE OF OUTPUTS FROM OUR SERVICE IS AT YOUR SOLE RISK AND YOU WILL NOT RELY ON OUTPUT AS A SOLE SOURCE OF TRUTH OR FACTUAL INFORMATION, OR AS A SUBSTITUTE FOR PROFESSIONAL ADVICE.

---

[31] Davies, D. (2025). *The unaccountability machine: Why big systems make terrible decisions* (Paperback edition). Profile Books.
[32] Bach, J. (2026). *Taking Testing Seriously: The rapid software testing approach*. JOHN WILEY & SONS, p. 267.
[33] https://openai.com/policies/row-terms-of-use (accessed 5/18/26)

It is certainly compatible with responsible work to use a service that specifically tells you that its vendor will not take any responsibility for its behavior. You can still be responsible as long as *you* supervise its output. Besides that, it's probably in the best *commercial* interests of the vendor, even if they accept no legal liability, to nevertheless engage responsible people to perform quality assurance on that service. (At least, until that company becomes a monopoly.)

Note that there is a new product liability directive in the EU[34] which prevents companies doing business with consumers from disclaiming all warranties such as OpenAI does, above.

Even if your company is offering a service for which they disclaim all warranties, you, as an employee still bear responsibility for your own work. You are responsible to your employer, at the very least. So, if you use AI to do your work, you must be ready to stand by it.

What's Not Here

This is a narrowly targeted set of principles that we believe are absolutely necessary for responsible use of AI, regardless of one's political views or variations of law across the world. Other treatments of this subject may also speak of such matters as inclusiveness, fairness, or sustainability.

We are not against discussing broader principles,[35] but we are focused here on the role human operators of AI play rather than the harm systems may do. We believe a sharp line must be drawn between people and machines, and that society must acknowledge that humans and machines are not interchangeable. There are people who think that the human/machine distinction is arbitrary and obsolete. We disagree strongly. The principles in this document are meant to frame our rebuttal to the dehumanization of employment.

Another out-of-scope question is *transparency*: whether a responsible person should disclose their use of AI to their clients. If you are using AI responsibly, we wouldn't consider that a necessity. People use all kinds of tools and props to get their work done, and it would be strange to think that we must disclose every little thing.

As a matter of business strategy (to preserve your reputation, for instance), you may want to be transparent about how you are using AI. We don't see that as a hard requirement. But, if you cannot assure that the quality of AI work is adequate, then you should warn your client that your service is unreliable.

---

[34] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32024L2853 (accessed 5/18/26)

[35] We think the IBM Responsible Technology Board is doing good work. Including their documentation of various AI risks: "Foundation Models: Opportunities, Risks and Mitigations." (2024). https://www.ibm.com/think/author/ibm-responsible-technology-board